Network-Based Delineation of Health Service Areas: A Comparative Analysis of Community Detection Algorithms

Diego Pinheiro[1][0000-0001-9300-7196], Ryan Hartman[3][0000-0001-8044-5461], Erick Romero[1][0000-0001-5456-3093], Ronaldo Menezes[2][0000-0002-6479-6429], and Martin Cadeiras[1][0000-0003-4545-2871]

[1] Department of Internal Medicine, University of California, Davis, USA
[2] Department of Computer Science, University of Exeter, UK
[3] Independent Researcher

Abstract. A Health Service Area (HSA) is a group of geographic regions served by similar health care facilities. The delineation of HSAs plays a pivotal role in the characterization of health care services available in an area, enabling better planning and regulation of health care services. Though Dartmouth HSAs have been the standard delineation for decades, previous work has recently shown an improved HSA delineation using a network-based approach, in which HSAs are the communities extracted by the Louvain algorithm in hospital-patient discharge networks. Given the known heterogeneity of communities extracted by different community detection algorithms, a comparative analysis of community detection algorithms for optimal HSA delineation is lacking. In this work, we compared HSA delineations produced by community detection algorithms using a large-scale dataset containing different types of hospital-patient discharges spanning a 7-year period in the USA. Our results replicated the heterogeneity among community detection algorithms found in previous works, the improved HSA delineation obtained by a network-based, and suggested that Infomap may be a more suitable community detection for HSA delineation since it finds a high number of HSAs with high localization index and a low network conductance.

Keywords: Hospital-Patient Discharge Networks · Community Detection Algorithms · Health Service Area · HSA Delineation

1 Introduction

A Health Service Area (HSA) is as a group of geographic regions in which residing patients most often receive healthcare services from similar health care facilities. It was first introduced in 1973 by Wennberg and Gittelsohn as a more meaningful unit of analysis for healthcare data than administrative geographic divisions such as counties [13]. By examining variations in expenditures among HSAs delineated in Vermont, the authors showed, for instance, that the expenditure per capita among HSAs varied from \$54 to \$162 and such variation, however, had no correlation with age-adjusted mortality.

In 1996, Wenneberg then proposed the *Dartmouth Atlas of Health Care in the United States* [12] which is the current standard HSA delineation in the US. Effectively, each Dartmouth-HSAs are delineated in three steps. First, each health care facility is assigned to its respective city/town. Then, each ZIP Code is assigned to the city/town of the health care facility from which residing patients receive most of their healthcare services. As a result, each Dartmouth-HSA is the group of ZIP Codes associated with the same city/town. Finally, enclave ZIP Codes, if any, are assigned the city/town of its adjacent ZIP Codes to unsure geographic contiguity.

Recently, Hu et al. [7] has proposed a network-based approach to HSA delineation in which a Hospital-Patient Discharge Network (HPDN) was built and the Louvain community detection algorithm was subsequently applied to find communities (i.e., HSAs) with the highest network modularity. In their HPDN, nodes represent distinct ZIP Codes and links represent the total number of discharges between the ZIP Codes of health care facilities and patient residences. Using claims-based hospital discharges in Florida, the authors demonstrated that Louvain-HSAs presented, for instance, a higher localization index than Dartmouth-HSAs, which is a measure of internal validity that quantifies the proportion of patients receiving services from health care facilities located within the same HSA in which they reside.

Yet, a comparative analysis that looks at the effectiveness of various community detection algorithms is still needed in order to attain optimal network-based HSA delineation. Such comparative analysis is needed given owing to the heterogeneity of communities extracted by different community detection algorithms [6]. Through a comparative analysis was previously provided for grouping hospitals [3], the underlying networks differ from HPDNs in two fundamental aspects: nodes were hospitals instead of geographical regions, and links were patients sharing between hospitals instead of the total number of hospital discharges.

In this paper, we conducted a comparative analysis of network-based community detection algorithms for HSAs delineated; we focus on four of the commonly used community detection algorithms, namely, Block Model [8], Infomap [10], Louvain [1], and Speaker-Listener Label Propagation Algorithm (SLPA) [14]. A claims-based patient-hospital discharge data was used; it contains a total of 124,970,471 discharges over a 7-year period in California, USA. Our results replicated the existent heterogeneity of communities extracted by different algorithms of community detection and reinforced the use of a network-based approach to HSA delineation. The results demonstrated, for instance, that Infomap was the most suitable algorithm because it was capable of delineating a high number of HSAs, while still presenting a high localization index and a low network conductance.

2 Methodology

A network-based delineation of Health Service Areas (Fig. 1) consists of (A) modeling claims-based patient-hospital discharge data as networks of ZIP Code Tabulation Areas (ZCTAs), (B) applying a diverse set of community detection algorithms, and (C) performing a comparative analysis of the extracted communities (i.e., HSAs) according to multiple quality metrics of HSA delineation.

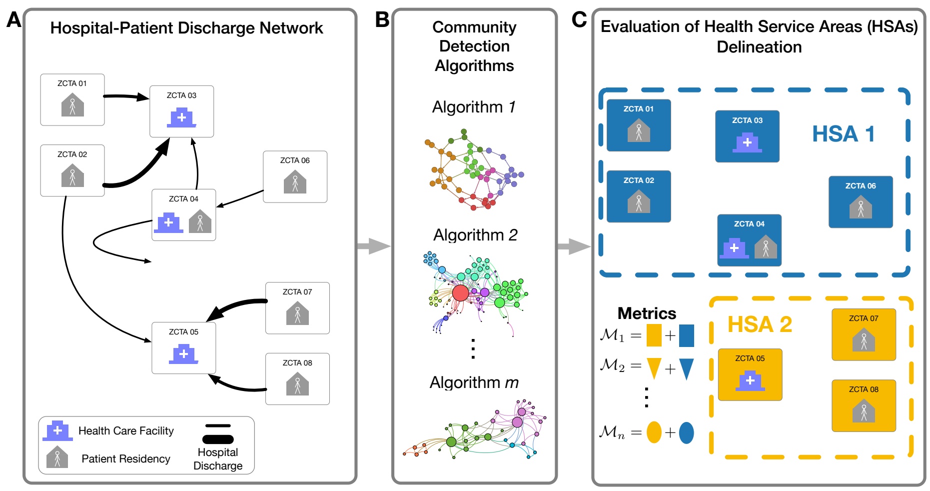

Fig. 1. Network-based delineation of Health Service Areas (HSAs). (**A**) Hospital discharge data is used to build Hospital-Patient Discharge Networks (HPDN) in which nodes represent ZIP Code Tabulation Areas (ZCTAs), and links represent the total number of discharges between the ZCTA locations of health care facilities and patient residencies. (**B**) Community detection algorithms are applied over HDPNs to delineate HSAs. (**C**) Delineated HSAs are evaluated according to multiple quality metrics of HSA delineation.

2.1 Hospital-Patient Discharge Networks

Hospital-Patient Discharge Networks (HPDNs) were built using claims-based hospital discharge data obtained from the California Health and Human Services Agency (CHHS) [2]. This dataset is publicly available and contains a total of 124, 970, 471 hospital-patient discharges of different types spanning a 7-year period from 2012 through 2018. The four discharge types are *Inpatient from ED*, *Inpatient*, *Ambulatory Surgery*, and *ED Only*. Each data point contains the type of discharge, the year, the name of the facility (e.g., Alameda Hospital, University of California Davis Medical Center). Also, it contains the 5-digit ZIP Codes (e.g., 94501, 95831) of both facility location and patient residency as well as the respective number of discharges between them. A discharge is the process by which patients undergo after the provision of the healthcare service when they leave the hospital. The healthcare services may require admission to the hospital for a overnight stay, inpatient, or may not require hospitalization, outpatient. A visit to the Emergent Department (ED), which is considered outpatient, may require further hospitalizations and thus become inpatient.

Hospital discharges to patient residency ZIP Codes other than those with 5-digits were excluded (2.6%); these were mainly discharges of patients from states other than California (e.g., ARIZONA, NEVADA (state), Other USA), from locations outside the US (e.g., OUTSIDE USA), from unknown locations (e.g., UNKNOWN), and from homeless population (e.g., HOMELESS). ZIP Codes are a collection of delivery routes maintained USA Postal Service and ZIP Code Tabulation Area (ZCTAs) are actual generalized areal representations maintained by the USA Census Bureau. Therefore, for each ZIP Code of both health care facility and patient residency, the corresponding ZCTA was obtained using the ZIP Code to ZCTA Crosswalk provided from the Uniform Data System Mapper [11].

A separate weighted and undirected HPDN network was built for each type of hospital discharge and for each year according to the methodology proposed by Hu et al. [7]. In each network, nodes are ZCTAs, links are the total number of discharges between the ZCTAs of health care facilities and patient residencies. The HPDN is an undirected network because a link w_{ij} encodes the total number of discharges *between* ZCTAs i and j without arbitrarily distinguishing whether the direction is due to a health care facility at i discharging patients residing at j or patients residing at i going to a facility at j for health care services. Overall, 28 Hospital-Patient Discharge Network (HPDN) were built, one for each combination of 4 discharge types and 7 years.

2.2 Community Detection

In a network-based HSA delineation, each HSA correspond to a distinct community extracted by a community detection algorithm from a hospital-patient discharge network (HPDN) [7]. While the existence of communities in real-world networks is agreed upon, there is no generally accepted definition of what a community is, or what the most appropriate way to find them is [5]. Some algorithms take a stronger approach to community detection by looking for cliques, which are a group of nodes for which there is a link between every pair of nodes [4]; other approaches just look for more densely connected subgraphs within the network such that a community is a subset of nodes within a network that are densely connected to each other when compared to the rest of the network.

The lack of a general definition of what a community should be is also mirrored in the existent heterogeneity of communities extracted by different community detection algorithms [6] and as such requires a comparison among community detection algorithms, particularly when the communities found are being used in important issues such as HSA delineation. Four commonly used algorithms were selected: Louvain Modularity, Infomap, Stochastic Block Model, and Speaker-listener Label Propagation. These algorithms were selected as they provide four very distinct approaches to community detection and have their implementations easily provided by their authors.

The *Louvain* algorithm [1] finds communities that maximizes the network modularity. This quantifies the extent to which the density of links within the found communities excessively surpasses that of what would be expected if links were placed at random. Trying all possible partitions is not computationally feasible, and Louvain modularity takes a heuristic approach by maximizing the local modularity of smaller communities that are only subsequently joined if such aggregation leads to an increased modularity. These smaller communities start as individual nodes and are iteratively joined together into greater communities until a single community containing the whole network is reached.

The *Stochastic Block Model* [8] uses a maximum likelihood estimator to infer the block structure of the network. This algorithm attempts to recover the hierarchical block structure of the network where each block represents a community. The block model used in this study is the degree-corrected variant given the weighted aspect of HPDNs and that previous work has shown that the such variant tends to perform better on empirical networks [9].

The *Infomap* algorithm [10] finds communities that maximizes the map equation instead of modularity. The map equation quantifies the length of the coding scheme necessary to communicate the sequence of movements of random walkers within the network. In essence, if a community structure exists, random walkers will tend to become trapped within these communities because movements within-communities are more likely than between-communities. Therefore, the coding scheme necessary to communicate the sequence of movements can be reduced by taking into account the community structure as every time a walker enters into a community, the community is identified, and a smaller community-specific coding scheme is used to quantify within-community movements.

The *Speaker-Listener Label Propagation Algorithm* (SLPA) [14] is a localized community detection algorithm based on the concept of label propagation. SLPA finds communities by initially assigning each node to a unique label. Nodes then iteratively changes their label to the label most often used by its neighbors. Initially, this label exchange rule promotes the formation of smaller consensus groups which will subsequently compete with one another for node members depending on the balance between their within-group and between-group interactions. In contrast to the other algorithm's, SLPA does not require prior information about the network, nor does it attempt to maximize any metric as a proxy for well-defined communities. Instead, it only relies on the network structure to identify the communities. While SLPA is able to find overlapping communities, the post processing threshold was set to $r = 0.5$ to ensure the extraction of non-overlapping communities and thus provide a better comparison to the other non overlapping algorithms used.

The application of the 4 community detection algorithms to the 28 HPDNs yielded a total of 112 HSA delineations. Though only a subset of HPDNs and HSA delineations are presented, all of the code, datasets, networks, and analysis are available on the Open Science Framework (OSF) repository of this project at https://doi.org/10.17605/OSF.IO/GW73Y.

2.3 Evaluation of Health Service Areas (HSA) Delineation

The quality of a HSA delineation can be evaluated according to multiple and often conflicting metrics [7,3]. The following four delineation metrics of a HSA c were used: the number of communities N_c, the localization index $LI(c)$, network conductance $C(c)$, and the total number of discharges $D(c)$.

In a network-based HSA delineation, the *total number of delineated HSAs* N_c is determined by the specific community detection algorithm used as it is the number of communities extracted. The N_c is an important metric since it determines the number of distinct meaningful units of analysis ultimately uncovered and, as in any community detection problem, trivial solutions such as 1 (one) and n, the total number of nodes, are generally undesired [6].

The *localization index* $LI(c)$ of a community c quantifies the proportion of patients seeking and receiving health care services from hospitals within the HSA in which they reside. In a community c with a higher $LI(c)$, residents mostly seek healthcare from hospital within the HSA c. Formally, the localization index $LI(c)$ of community c can be defined as

$$LI(c) = \frac{D(c,c)}{D(c)} \quad , \tag{1}$$

in which $D(r, s)$ is the number of discharges of patients residing at ZCTAs within community r that are discharged from hospitals at ZCTAs within community s, and $D_c = \sum_s^{N_c} D(c, s)$ is the total number of discharges from patients living within community c.

The network conductance $C(c)$ of a community c is a network-based measure which quantifies the extent to which c is a well-formed community by comparing the total links running within community c relative to links running from c to other communities. Conductance is based on the degree-based definition of a community [5]. Formally, the conductance $C(c)$ of a community c can be calculated as

$$C(c) = \frac{w^{ext}(c)}{w(c)} \quad , \tag{2}$$

in which, $w(c) = \sum_{i \in C, j} W_{ij}$ is the total strength of links originating within community c, $w^{ext}(c) = \sum_{i \in C, j \notin C} W_{ij}$ is the external strength, and W_{ij} is the strength of the link connecting nodes i and j.

Aside from the aforementioned metrics, the total number of discharges $D(c)$ from patients living in one of the ZCTAs found within community c is also calculated. Ideally, a community detection algorithm would maximize the number of HSAs where each has a high localization index and a low conductance. Yet, as the total number of communities increases from one to n, the typical value of localization index decreases from 1 to 0 and the typical value of network conductance increases from 0 to 1.

To provide a reliable estimate for all communities found by a single community detection algorithm, $B = 1,000$ bootstrap samples with replacement were draw from the distribution of each metric to calculate a mean value for the localization index $\langle LI(c)\rangle$, network conductance $\langle C(c)\rangle$, and total number of discharges $\langle D(c)\rangle$. The standard deviation was also provided for each of the aforementioned estimators.

3 Results and Discussion

The network statistics of each individual HPDN varied among discharge types and over the years (Table 1). Considering the year of 2012, for instance, *Inpatient from ED* and *ED Only* HPDNs had comparable total number of nodes n. Their number of links m were $49,000$ and $127,000$, respectively, suggesting that a *ED Only* HPDN has a higher number of distinct ZCTA pairs for which hospital discharges occurred between hospital locations and patient residencies. Also, their the total network link strength w were 1.6 million and 9.2 million, respectively, and their network density ρ were 0.0345 and 0.0883, respectively,suggesting that the healthcare service underlying a *ED Only* HPDN has a higher demand and is more dense. Over the years, the average shortest path length l and clustering coefficient c of HPDNs slightly decreased.

Overall, HSA delineation results (Fig. 2 and Table 2) have reinforced the superiority of network-based HSA delineation as well as confirmed the heterogeneity among communities extracted by different community detection algorithms even in the same dataset. Such results are not only consistent with community detection comparisons from previous works [6], but also advocate for a further comparison among community detection algorithms.

The comparison of HSA delineations (Fig. 2) involves the evaluation of conflicting metrics such as a higher number of communities and localization index. Considering the discharge type *ED Only* and year 2018 (Table 2), for instance, the difference between the number Louvain-HSAs and SLPA-HSAs was 4-fold, with 24 Louvain-HSAs and 126 SLPA-HSAs. This fewer number of Louvain-HSAs using hospitals discharge in California is also consistent with the fewer number of Louvain-HSAs in the previous work from Hu et al. [7] using hospital discharges in Florida. Though the localization index of Louvain-HSAs (.90) was 7% higher than that of Infomap-HSAs (.84), 52 more Infomap-HSAs were delineated, representing a 2-fold increase.

By examining their geographical patterns (Fig. 3), the respective geographical areas of Louvain-HSAs correspond to a wider and more discontinuous geographical areas than those of Infomap-HSAs. Yet, Block Model-HSAs appear to be the poorest HSAs delineated not only because they are the fewest number of HSAs delineated, which corresponds to a wider and more discontinuous area, but also as their localization index is lower than .5 which raises concerns about the internal validity of the HSAs delineated. Lastly, Block Model-HSAs presented the highest variability regarding their localization index, conductance, and total number of discharges. Using the nested version of Block Model has not

Table 1. Statistics of Hospital-Patient Discharge Networks (HPDNs). For each type of discharge and for each year, the following network measures were quantified: the total number of nodes (n), the total number of links (m), the total weight (w), the network density (ρ), the average shortest path length (l), as well as the clustering coefficient (c). Other types of discharges and network metrics are available on the Open Science Framework (OSF) repository of this project at `https://doi.org/10.17605/OSF.IO/GW73Y`.

Type of Discharge	Year	n	m	w	ρ	l	c
Inpatient from ED	2012	1.69×10^3	4.92×10^4	1.63×10^6	3.45×10^{-2}	2.81	1.19×10^{-3}
	2013	1.69×10^3	4.97×10^4	1.63×10^6	3.49×10^{-2}	2.72	1.12×10^{-3}
	2014	1.69×10^3	5.05×10^4	1.64×10^6	3.56×10^{-2}	2.71	1.02×10^{-3}
	2015	1.68×10^3	5.25×10^4	1.72×10^6	3.70×10^{-2}	2.89	9.74×10^{-4}
	2016	1.75×10^3	5.37×10^4	1.74×10^6	3.50×10^{-2}	2.70	9.53×10^{-4}
	2017	1.75×10^3	5.38×10^4	1.75×10^6	3.52×10^{-2}	2.67	9.84×10^{-4}
	2018	1.75×10^3	5.41×10^4	1.75×10^6	3.54×10^{-2}	2.62	9.31×10^{-4}
ED Only	2012	1.69×10^3	1.27×10^5	9.25×10^6	8.83×10^{-2}	2.15	2.14×10^{-4}
	2013	1.69×10^3	1.28×10^5	9.65×10^6	8.95×10^{-2}	2.18	2.07×10^{-4}
	2014	1.69×10^3	1.33×10^5	1.03×10^7	9.25×10^{-2}	2.20	1.90×10^{-4}
	2015	1.69×10^3	1.39×10^5	1.12×10^7	9.66×10^{-2}	2.16	1.78×10^{-4}
	2016	1.76×10^3	1.42×10^5	1.15×10^7	9.15×10^{-2}	2.15	1.81×10^{-4}
	2017	1.76×10^3	1.43×10^5	1.17×10^7	9.25×10^{-2}	2.15	1.88×10^{-4}
	2018	1.76×10^3	1.42×10^5	1.14×10^7	9.17×10^{-2}	2.16	1.96×10^{-4}

changed these results. Conversely, SLPA-HSAs are the highest number of HSAs but their localization index is typically less than .7, which is higher than that of Block Model-HSAs but still raises concerns regarding the internal validity of SLPA-HSAs.

The hospital discharge data and the Hospital-Patient Discharge Networks (HPDNs) inherently represent a flow between hospitals and patients and the apparent superior results achieved by Infomap may be related to the fact that, instead of modularity, the Infomap optimizes the map equation and thus takes into account the local flows emerging from the movements of random walkers trapped within the HPDN communities. Interestingly, Infomap-HSAs presented improved localization index and conductance over time.

4 Conclusions

Health Service Areas (HSAs) are the meaningful units of analysis for improving the scientific basis of both clinical practice and policy decision making in the delivery of health care. The optimal delineation of HSAs is necessary to create

Table 2. Comparison of HSAs delineated using the community detection algorithms Block Model, Infomap, Louvain, and SLPA in terms of in terms of the number of communities (n_c) as well as the typical values of the localization index $\langle li \rangle$, network conductance $\langle c \rangle$, and total number of discharges $\langle d \rangle$. The discharge types presented are *Inpatient from ED* and *ED Only* for the years of 2012 and 2018. The other discharge types and years are available on the Open Science Framework (OSF) repository of this project at `https://doi.org/10.17605/OSF.IO/GW73Y`.

Type of Discharge	Year	Community Detection	$N(c)$	$\langle LI(c) \rangle$	$\langle C(c) \rangle$	$\langle D(c) \rangle$
Inpatient from ED	2012	BLOCK MODEL	35	0.47	0.82	51,450
		INFOMAP	70	0.77	0.18	25,539
		LOUVAIN	20	0.86	0.13	89,235
		SLPA	110	0.69	0.25	16,392
	2018	BLOCK MODEL	35	0.47	0.74	54,090
		INFOMAP	62	0.80	0.15	30,397
		LOUVAIN	15	0.87	0.13	125,833
		SLPA	111	0.65	0.28	16,830
ED Only	2012	BLOCK MODEL	33	0.45	0.84	311,009
		INFOMAP	90	0.77	0.22	114,117
		LOUVAIN	24	0.92	0.09	430,396
		SLPA	139	0.70	0.28	74,049
	2018	BLOCK MODEL	34	0.45	0.83	359,908
		INFOMAP	76	0.84	0.15	160,869
		LOUVAIN	24	0.90	0.09	512,044
		SLPA	126	0.73	0.25	97,810

not only more meaningful units of analysis, but also to characterize medical practices with greater accountability regarding their respective community needs and shared care practices.

As the delineation approach of HSAs shift from the Dartmouth towards a network-based, further work will be needed to establish a comprehensive methodology for network-based HSA delineation, which should include (i) a broader set of community detection algorithms, (ii) hospital discharge data from states other than California, and integration with other healthcare datasets of utilization, expenditures, and outcomes.

References

1. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics-Theory and Experiment*, 2008(10), October 2008.

2. CHHS. California health and human services agency. `https://data.chhs.ca.gov`. Accessed: 2019-11-15.
3. Jordan Everson, John M Hollingsworth, and Julia Adler-Milstein. Comparing methods of grouping hospitals. *Health Services Research*, 24(4):333–9, June 2019.
4. Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, February 2010.
5. Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, November 2016.
6. Ryan Hartman, Josemar Faustino, Diego Pinheiro, and Ronaldo Menezes. Assessing the suitability of network community detection to available meta-data using rank stability. *WI*, pages 162–169, 2017.
7. Yujie Hu, Fahui Wang, and Imam M Xierali. Automated Delineation of Hospital Service Areas and Hospital Referral Regions by Modularity Optimization. *Health Services Research*, 53(1):236–255, February 2018.
8. Tiago P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1), March 2014.
9. Tiago P. Peixoto. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, 5(1), March 2015.
10. M. Rosvall and C. T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, January 2008.
11. UDS. Uniform data system mapper. zip code to zcta crosswalk. `https://www.udsmapper.org/zcta-crosswalk.cfm`. Accessed: 2019-11-15.
12. John Wennberg. *The Dartmouth Atlas of Health Care*. American Hospital Association, Chicago, IL, 1996.
13. John Wennberg and Alan Gittelsohn. Small Area Variations in Health Care Delivery: A population-based health information system can guide planning and regulatory decision-making. *Science*, 182(4117):1102–1108, December 1973.
14. Jierui Xie, Boleslaw K. Szymanski, and Xiaoming Liu. SLPA: uncovering overlapping communities in social networks via A speaker-listener interaction dynamic process. *CoRR*, abs/1109.5720, 2011.

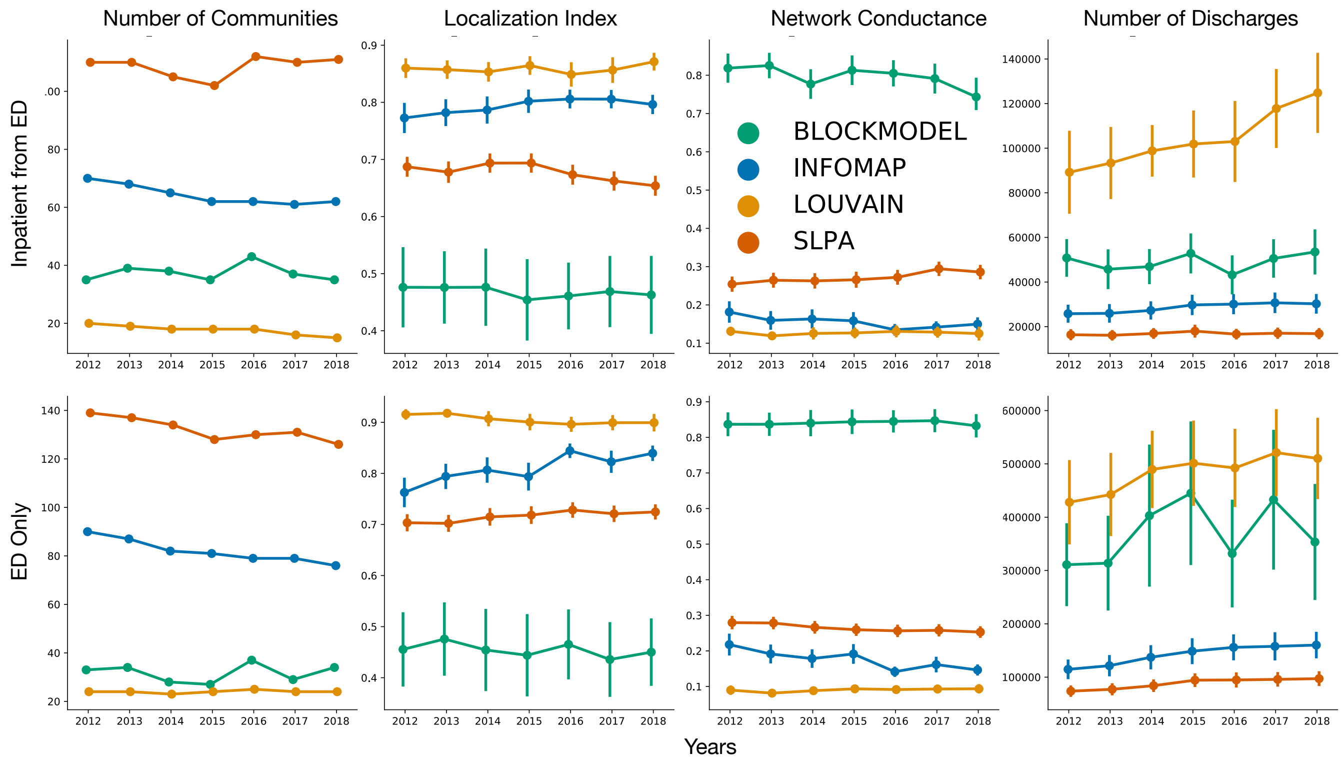

Fig. 2. Comparison of HSAs delineated using Block Model, Infomap, Louvain, and SLPA community detection algorithms. The HSA delineations were compared in terms of the number of communities (N_C), localization index ($LI(c)$), network conductance ($C_(c)$), and total number of discharges ($N_D(i,j)$). The discharge types presented are *Inpatient from ED* (top) and *ED Only* (bottom). All results are available on the Open Science Framework (OSF) repository of this project at `https://doi.org/10.17605/OSF.IO/GW73Y`.

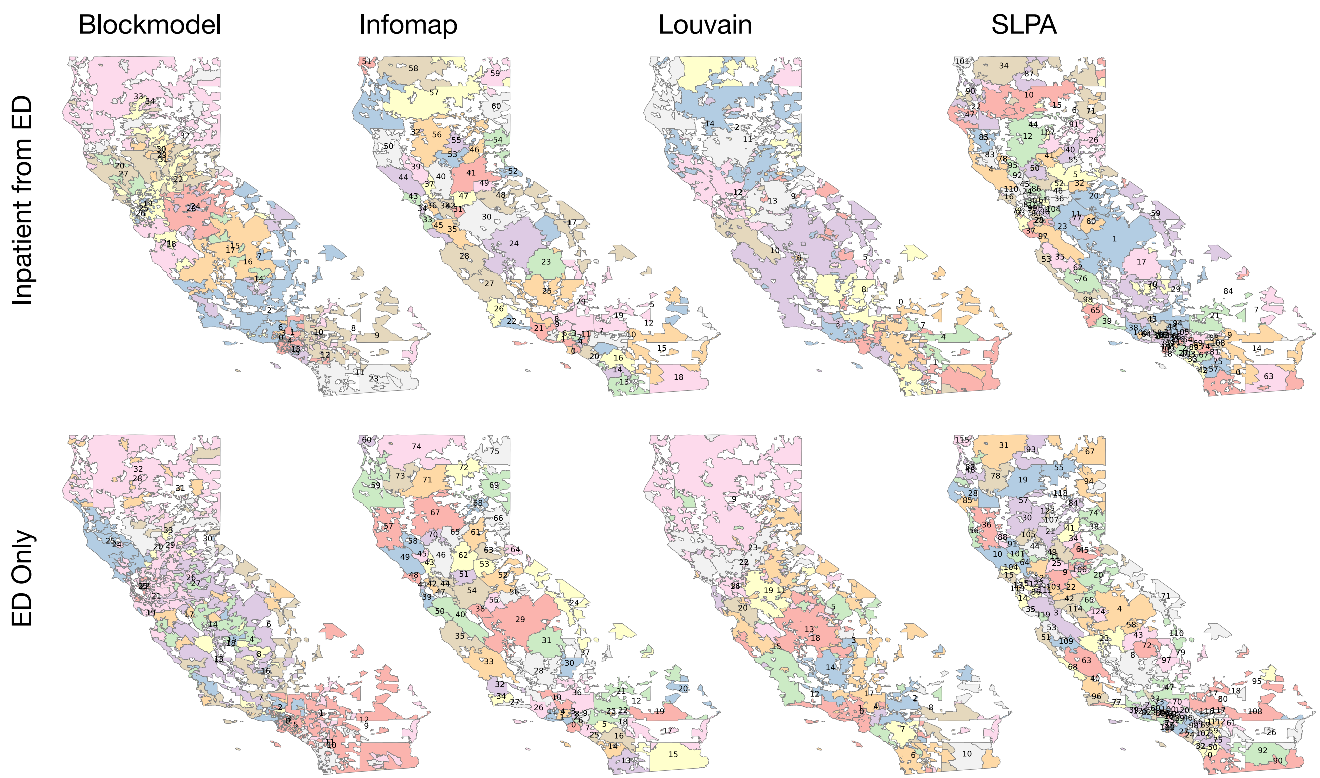

Fig. 3. Maps of HSA delineated using the community detection algorithms Block Model, Infomap, Louvain, and SLPA for discharge types *Inpatient from ED* (top) *ED Only* (bottom) in 2018. Each HSA is displayed as a geographical boundary aggregating one or more ZCTAs along with its respective Id. HSAs were then colored using a color blind palette with 9 distinct colors. All maps are available on the Open Science Framework (OSF) repository of this project at `https://doi.org/10.17605/OSF.IO/GW73Y`